“Integrated Genomic Characterization of Papillary Thyroid Carcinoma”

Supplementary Information: Integrated Analysis and Interactive Exploration using Regulome Explorer

Workgroup leader: Lisa Iype (liype@systemsbiology.org)
Contributors: Sheila M. Reynolds (sheila.reynolds@systemsbiology.org), Richard Kreisberg (dick.kreisberg@systemsbiology.org), Roger Kramer, Michael Miller, Brady Bernard, Vesteinn Thorsson, Ilya Shmulevich, and *Wei Zhang*

To gain greater insight into the development and progression of papillary thyroid carcinoma, we have integrated all of the data types produced by TCGA and described in this paper into a single “feature matrix”. From this single heterogeneous dataset, significant pairwise associations have been inferred using statistical analysis and can be visually explored in a genomic context using Regulome Explorer, an interactive web application (http://explorer.cancerregulome.org). In addition to associations that are inferred directly from the TCGA data, additional sources of information and tools are integrated into the visualization for more extensive exploration (e.g., NCBI Gene, miRBase, the UCSC Genome Browser, etc).

Feature Matrix Construction. A feature matrix was constructed using all available clinical, sample, and molecular data for 496 unique patient/tumour samples. The clinical information includes features such as age, gender, tumour size; while the sample information includes features derived from molecular data such as single-platform cluster assignments and mutation rates. The molecular data includes mRNA and microRNA expression levels (Illumina HiSeq data), protein levels (RPPA data), copy number alterations (derived from segmented Affymetrix SNP data as well as GISTIC regions of interest and arm-level values), DNA methylation levels (Illumina Infinium Methylation 450k array), and somatic mutations. For mRNA expression data, gene level RPKM values from RNA-seq were log2 transformed, and filtered to remove low-variability genes (bottom 25% removed, based on interdecile range). For miRNA expression data, the summed and normalized microRNA quantification files were log2 transformed, and filtered to remove low-variability microRNAs. (An initial filter removed any microRNA not observed in at least 9 samples, and a second filter removed the bottom 25% by interdecile range.) For methylation data, probes which are common between the 27k and the 450k platforms were used, and then the probes were filtered to remove the bottom 25% based on interdecile range. For somatic mutations, several binary mutation features indicating the presence or absence of a mutation in each sample were generated. Mutation types considered were synonymous, missense, nonsense and frameshift. Protein domains (InterPro) including any of these mutation types were annotated as such, with nonsense and frameshift annotations being propagated to all subsequent protein domains.

Pairwise Statistical Significance. Statistical association among the diverse data types in this study was evaluated by comparing pairs of features in the feature matrix. Hypothesis testing was performed by testing against null models for absence of association, yielding a *p*-value. *P*-values for the association between and among clinical and molecular data types were computed according to the nature of the data levels for each pair: categorical vs. categorical (Chi-square test or Fisher’s exact test in the case of a 2x2 table); categorical vs. continuous (Kruskal-Wallis test) or continuous vs. continuous (probability of a given Spearman correlation value). Ranked data values were used in each case. To account for multiple-testing bias, the *p*-value was adjusted using the Bonferroni correction.

Exploring significant associations between molecular features. Regulome Explorer allows the user to interactively explore significant associations between all of the various types of molecular features. One example is illustrated in Figure 1 which shows the top 100 significant associations between microRNA features and gene expression (mRNA) features where there is a negative correlation relationship. MicroRNAs hsa-mir-21-5p at chromosome 17q23 and hsa-mir-146b-5p at 10q24 are labeled. An arc indicates an association between a microRNA and a gene. Hovering over a single arc allows the user to see additional feature information, and clicking on the arc produces a scatterplot of the underlying data.

Figure S76

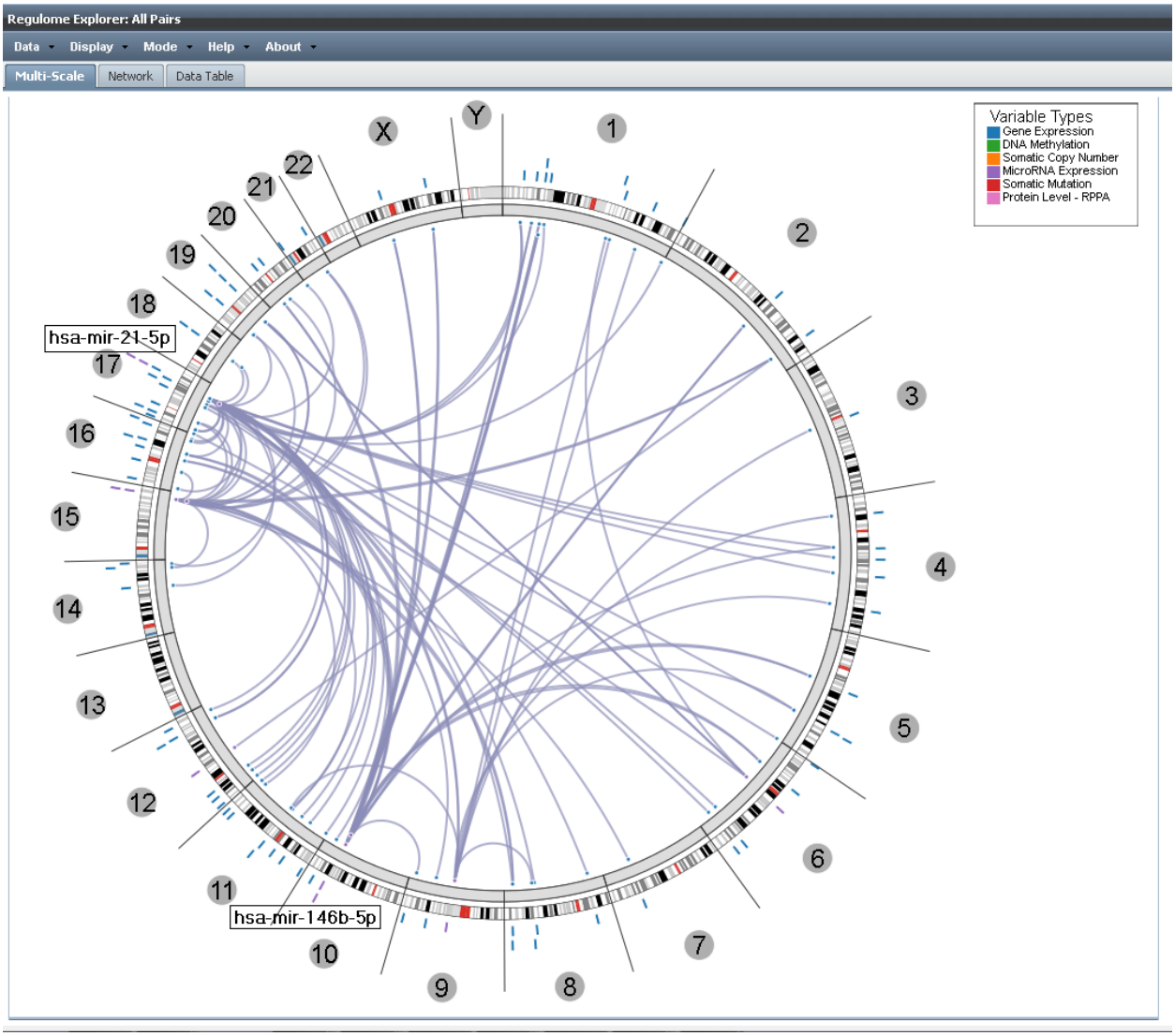

Figure S76. Associations between molecular features. The top 100 negative associations between microRNA and mRNA expression are shown. Statistically significant associations between features with genomic coordinates are indicated by arcs connecting pairs of dots which represent the features. The purple tick marks represent miRNAs while the blue tick marks represent mRNAs. hsa-mir-21-5p on chromosome 17 and hsa-mir-146b-5p on chromosome 10 are labeled.

Exploring significant associations with numeric scores. In addition to exploring associations between molecular features, it is also useful to explore the associations between molecular features and numeric features such as the *BRAF*^{V600E}-*RAS* score. In this case, statistically significant associations between molecular features (with genomic coordinates) and a numeric feature are shown as dots on a circular graph with a radial axis representing negative log10 (p-value). Figure 2 shows the top 100 miRNA associations for *BRAF*^{V600E}-*RAS* score.

Figure S77

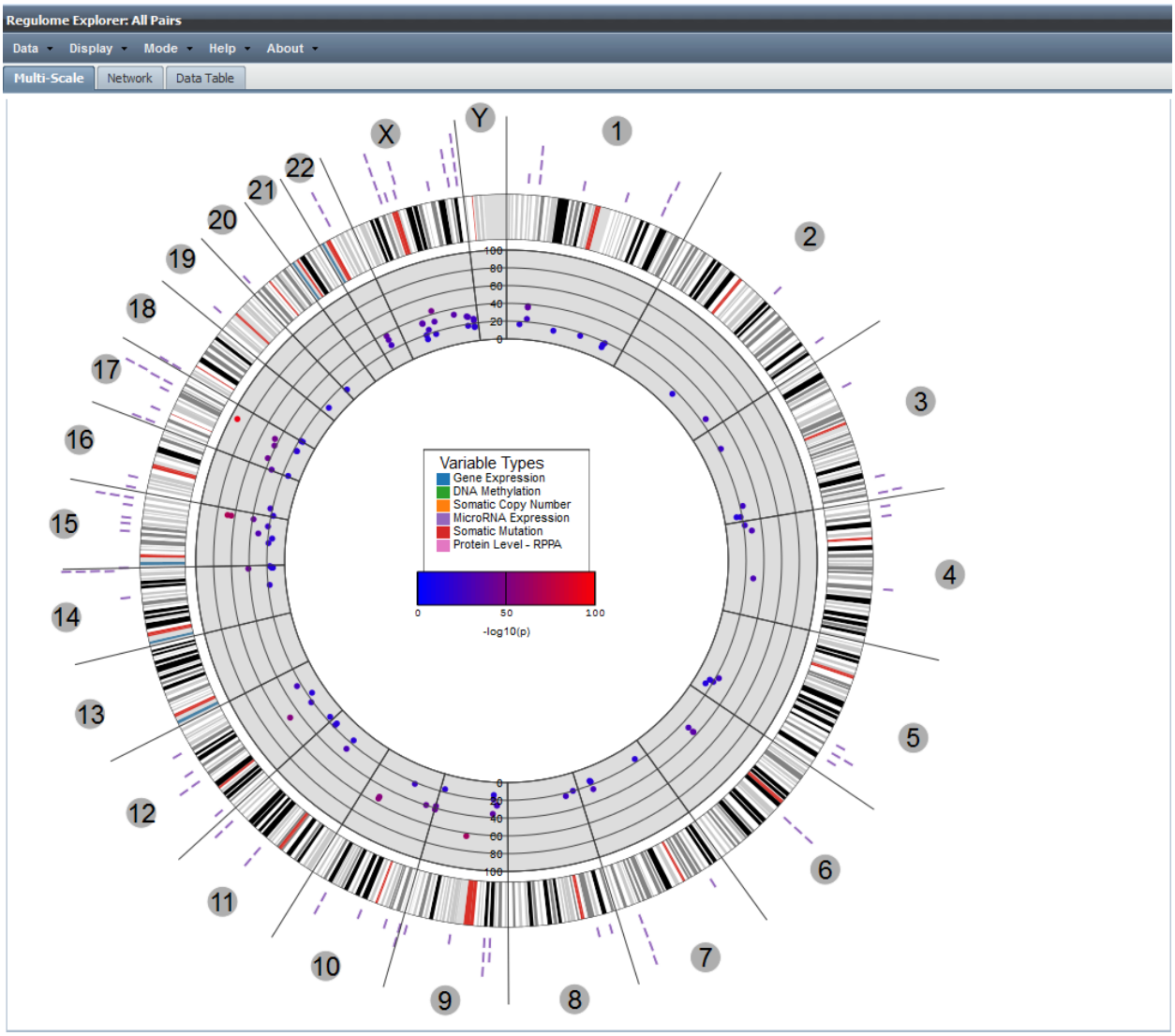

Figure 77. The top 100 miRNA associations with *BRAF*^{V600E}-*RAS* score are shown.

Exploring significant associations with a subtype. Regulome Explorer can be used to explore the associations between molecular features and categorical features such as a histological subtype or a subset of the tumour samples based on some other prior analysis such as expression-based clusters. Figure 3 shows the top 100 molecular associations for mRNA cluster 5.

Figure S78

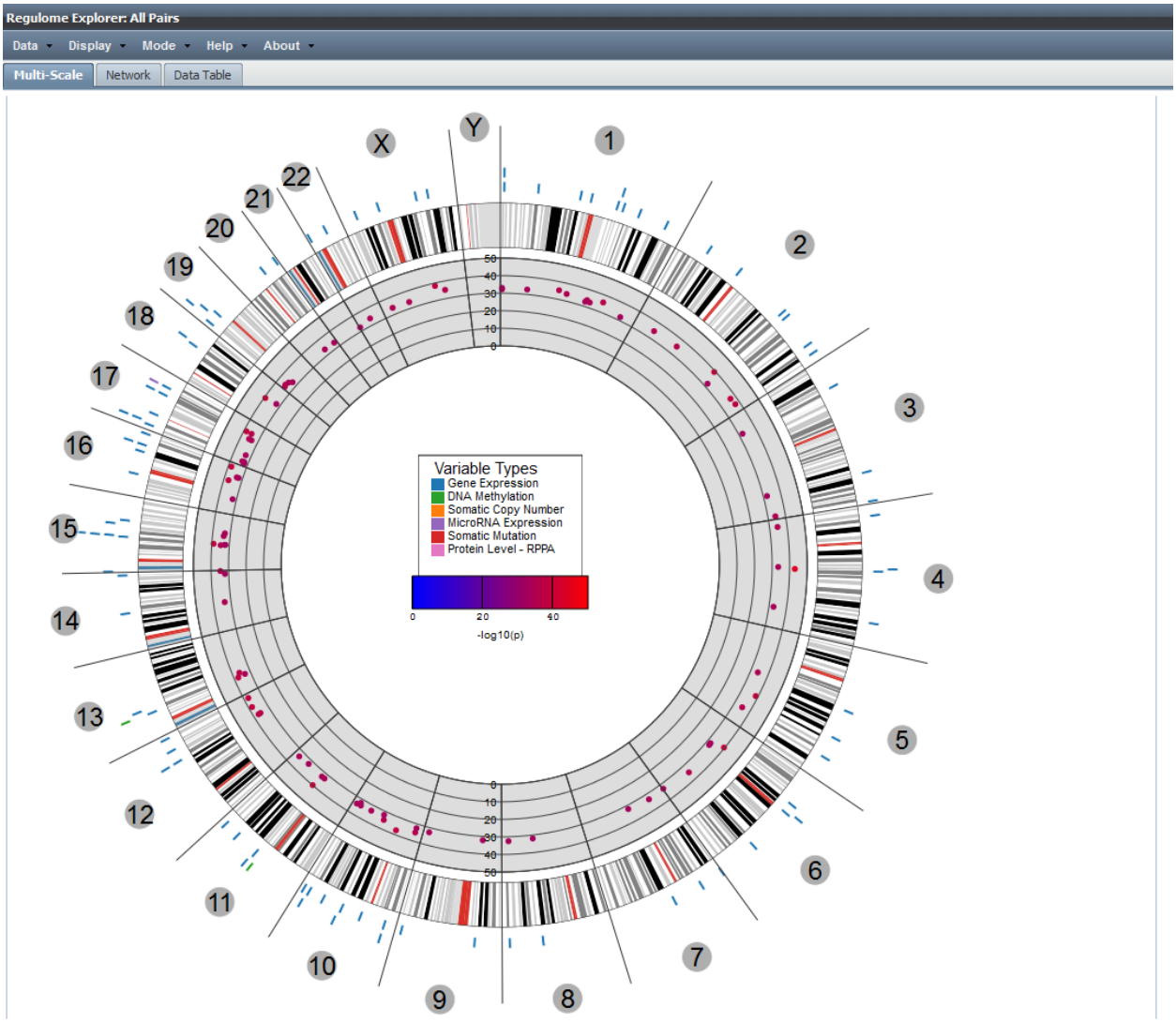

Figure 78. Associations between molecular and categorical features. The top 100 molecular associations with mRNA cluster 5 are shown.